QSPR Model for Regulatory Purpose: from Development to Integration into the QSAR Toolbox

Guillaume Fayet, Patricia Rotureau*

Institut National de l'Environnement Industriel et des Risques (INERIS), Parc Technologique Alata, BP2, 60550 Verneuil-en-Halatte, France
patricia.rotureau@ineris.fr

Quantitative Structure-Property Relationships (QSPR) are predictive methods of macroscopic properties of substances based on their only molecular structures. If these methods were initially mainly devoted to biology and toxicology, they are nowadays increasingly used for the prediction of physico-chemical properties.
In the framework of the European REACH regulation (2006), their developments are encouraged as alternative to experimental tests. Indeed, this regulation requires the evaluation of the physico-chemical properties of a large number of existing substances (143 000 pre-registered substances in 2008) in order to allow their use before 2018. Due to the number of substances and properties, the timing, the economic costs, the feasibility at the R&D level and the risks for the manipulator, in particular for the characterization of the dangerous physico-chemical properties (explosibility, flammability), the complete gathering of the data solely based on experimental measurements is not realistic. Thus, the use of such alternative predictive methods was recommended by REACH for the evaluation of the properties of substances.
In this context, INERIS developed QSPR models for the prediction of hazardous physico-chemical properties of chemical substances like explosibility of nitro compounds, thermal stability of organic peroxides or flammability of amines. To this end, quantum chemical tools allowed calculating relevant molecular descriptors (notably issued from Conceptual DFT) and to evidence subjacent chemical mechanisms. These models were developed according to the five validation principles of QSPR models proposed by OECD for regulatory use. During the French PREDIMOL (molecular modeling prediction of physico-chemical properties of products) project (2015) funded by ANR (National Research Agency), it was demonstrated that molecular modeling, notably the use of QSPR models, was a credible alternative approach to experimental characterization to access, in a reliable and fast manner, physico-chemical properties of substances required by REACH (annexes VII and IX).
This paper presents an overview of existing validated QSPR models for hazardous physico-chemical properties including those developed recently by INERIS. Then, the different ways towards the acceptance and use of QSPR in a regulatory assessment of chemicals are discussed and exemplified with the first QSPR model implemented into the QSAR Toolbox of ECHA and OECD (2015) (in its version 3.3 of December 2014) for the prediction of a hazardous physico-chemical property. This model is a simple multilinear model (based only on constitutional descriptors) dedicated to the prediction of one property of explosibility which is the impact sensitivity of nitroaliphatics (Prana et al. 2012).

1. Overview of existing QSPR models for hazardous physico-chemical properties

1.1 Principle of the QSPR method

The QSPR method is based on the principle that molecules with similar structures have similar properties. The molecular structure is represented by a series of descriptors that can be mathematically connected to experimental properties by a QSPR model. So, such model will have the following form:

Property =f(descriptors) (1)

A large number of descriptors (constitutional, topological, geometric and quantum chemical) can be calculated to describe the structure of molecules (Karelson, 2000). In this paper, each molecule was characterized by

more than 400 descriptors calculated in Codessa software (2002) based on geometric structures optimized at the PBE0/6-31+G(d,p) level of DFT (Density Functional theory) using Gaussian09 (2009). These descriptors include quantum chemical ones and in particular those issued from the conceptual DFT but also additional descriptors extracted from the examination of chemical structures (for example, the number of specific fragments such as n_{CNC} (the number of CNC fragments) or the oxygen balance).
Many statistical tools can be used to develop QSPR models (multi-linear regression (MLR), genetic algorithm, neural network, principal component analysis (PCA), decision tree, etc...). In this paper, the presented MLR models were obtained using the Best Multi Linear Regression (BMLR) technique (Karelson, 2000), a stepwise approach, as implemented in Codessa software (2002), to choose the final model as the best compromise between correlation and number of descriptors.
Within the context of REACH, the development of QSPR models is encouraged providing that they respect the 5 following principles for the validation of QSPR models drawn up by OECD (2007):

1. the endpoint must be defined, including experimental protocols;
2. the algorithm of the model must be unambiguous, with full definition of the descriptor;
3. its applicability domain is defined;
4. appropriate measures of goodness-of-fit, robustness and predictive power are provided;
5. if possible, a mechanistic interpretation is proposed.

The fourth OECD principle requires appropriate measures of performances by a series of internal and external validations (Gramatica, 2007). The goodness-of-fit is measured by the determination coefficient R^2 and the root mean square error (RMSE) between predicted and experimental values. For robustness and chance, the Q^2 coefficients issued from leave-one-out (LOO) and leave-many-out (LMO) cross-validations are computed. Y-randomization aims to ensure against chance correlation by checking that models issues from erroneous randomized property values give low mean values and standards deviation in R^2 (denoted R^2_{YS} and SD_{YS}, respectively). Then the predictive power of models is evaluated on an external validation set of compounds by a series of coefficients (Chirico and Gramatica, 2012): R^2_{ext}, $RMSE_{ext}$, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} and CCC.
The applicability domain (AD) required by the third OECD principle i.e. the domain in which predictions can be considered as accurate, was defined by the molecules of the training set. Two methods were used. Euclidean distance method available in Ambit discovery software (Jeliazkova, 2007) was used with a 95% threshold, i.e. the domain was calculated to contain 95% of the molecules of the training set. A simplest method to define AD was used consisting in defining ranges of values of properties and descriptors (for those included into the model) in the training set. Then, the performances inside the AD were also calculated based only on the molecules of the validation set that belonged to this domain using coefficients previously presented (R^2_{in} and $RMSE_{in}$).

1.2 Existing QSPR models for hazardous physico-chemical properties and overview of models developed by INERIS

QSPR models represent powerful tools already used for biological, toxicological, pharmaceutical and physico-chemical applications (Katritzky, 2010) but also for hazardous substances linked to the REACH regulation (Dearden, 2013). This last review indicated that if a lot of QSPR models have been developed for decades considering physico-chemical properties, few of them were developed according to OECD validation principles allowing their used in regulations related to chemicals. Indeed, if we consider hazardous physico-chemical properties, only one part of them are considered with validated QSPR models such as explosive substances, flammable gases and liquids and organic peroxides. For some others like oxidizing properties, no QSPR models exists due to the lack of available data and/or the nature of the test that give qualitative (yes/no) result.
Upon the type of descriptors and the type of data mining tools, INERIS developed more than forty QSPR models dedicated to various substances and hazardous properties (see Figure 1 and Table 1 for performances). Substances and properties were chosen according to a detailed examination of the literature considering existing QSPR models for REACH regulation (that has put in evidence for example, a lack of data and models for organic peroxides) and considering that INERIS is an expert in the experimental characterization of explosives. As these properties are used in regulatory frameworks (e.g. for the transport of dangerous goods (2011) or for the REACH registration of chemicals), models were developed in agreement with the OECD principles of validation. All these models were published in international papers and are easily accessible. It's interesting to note that for one property, various models were developed considering all types of descriptors (including quantum chemical ones) or only simple descriptors (such as constitutional and topologic) depending on the expected use of the model. Indeed, one can prefer to use a fast and easy model for screening purpose without any expert knowledge needed or to use more complex model needed prior quantum chemical calculations (sometimes more accurate than other ones) to better understand involved mechanisms. These models are complementary one to each other and can be used in a prediction process into a consensus approach.

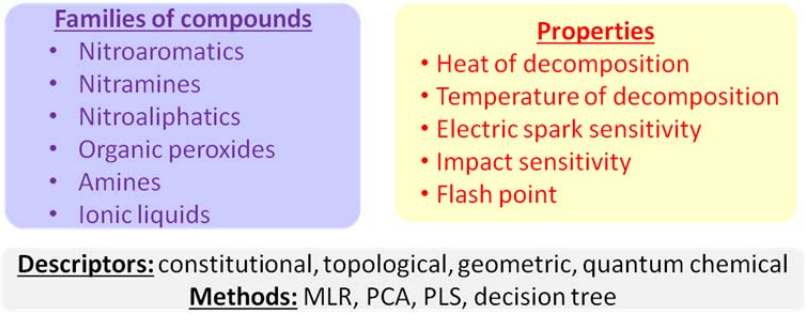

Figure 1: Scope of QSPR models developed at INERIS

It's worth noting that most of models were developed for pure compounds. Only recently, INERIS proposed predictions for the properties of mixtures, in particular for the flash points of organic liquid mixtures (due to the important number of data available in literature for such development). A first approach consisted in including QSPR models to provide predictions on pure compounds in existing mixing rules (Gaudin et al., 2014) and gave good predictive performances (with errors about 5K on binary mixtures). Another strategy was also developed by Gaudin et al. 2015 consisting in defining mixture descriptors by applying mixing formula on molecular descriptors to achieve mixture QSPR models. Evaluated on an external validation set of data, this model gave promising performances with a mean absolute error of 10.3°C.
At last, to access and validate robust experimental data (in terms of number and uncertainty of data), new experimental databases must be built, as done, for example, for the thermal stability of organic peroxides in the PREDIMOL project.

Table 1: Performances of INERIS QSPR models for hazardous physico-chemical properties

Explosive substances		
	Temperature and heat of decomposition	Recent validated models have been developed for nitro compounds with errors lower than 20% on heats of decomposition (Fayet et al. 2011).
	Impact sensitivity	Few validated models exist for nitro compounds with errors about 0.2-0.25 (log) (Prana et al., 2012; Fayet et al., 2014)
Flammable liquids		
	Flash point	A validated model applicable to amines exists with R^2 of 0.91 in external validation (Fayet et al., 2013). First predictive approaches have been recently proposed for mixtures with errors about 4°C (Gaudin et al. 2014, 2015).
Organic peroxides		
	Temperature and heat of decomposition	Validated models have been recently developed with R^2 of 0.82 and 0.90 (in external validation) for the heat and temperature of decomposition, respectively (Prana et al. 2014).

Finally, predicted data obtained from QSPR models can complete databases of experimental data. They can be used for screening molecules in R&D, even when substances have not been yet synthesized or in substitution purposes while reducing costs associated to tests. To encourage further their use in regulatory context such as process safety analyses or for the classification and registration of chemicals according to TDG, CLP or REACH (e.g. to fulfil safety dossiers or safety data sheets), efforts and dissemination of results have to be made among industry and regulatory bodies.

2. QSPR models as alternative method for regulatory purpose

2.1 Regulatory use of QSPR models

QSPR are identified among relevant methods that can be used as an alternative to experimental testing as stated in Annex XI of REACH: "*Results obtained from valid qualitative or quantitative structure-activity*

relationship models ((Q) SARs) may indicate the presence or absence of a certain dangerous property". Moreover, guidance on the use of grouping and QSARs in REACH (ECHA, 2008) detailed possible uses of such models like quantitative structure-property relationships:

"(a) provide information for use in priority setting procedures;
(b) guide the experimental design of an experimental test or testing strategy;
(c) improve the evaluation of existing test data;
(d) provide mechanistic information (which could be used, for example, to support the grouping of chemicals into categories);
(e) fill a data gap needed for hazard and risk assessment;
(f) fill a data gap needed for classification and labelling;
(g) fill a data gap needed for PBT or vPvB assessment."

As explained by Worth (2010), to allow application of QSPR models for regulatory purpose, the validity, applicability and adequacy of the model have to be demonstrated. The different steps are summarized in brief in the following. Concerning the validation of QSPR models, a first general agreement is based on the fact that the model should be scientifically valid or validated. For the validity, five validation principles proposed by OECD (2007) should be satisfied. Nevertheless, no performance criteria for the regulatory acceptance of QSPR were provided. The appropriate documentation used to demonstrate that the QSPR model fulfill all requirements are compiled into a QSAR Model Reporting Format (QMRF) file which is structured according to these principles. Some validated models and their associated QMRF files can be found in the online QSAR database of the Joint Research Center (2015), in which more than 90 models are included after validation by an expert comity, or in the OECD/ECHA QSAR Toolbox (2015), a free predictive platform available online in which QSAR and QSPR models are implemented (also after acceptance by an expert comity on the scientific validity and the technical feasibility of implementation of the model in the platform).

Moreover, whatever the quality and validation of the model, it also has to be relevant and correctly used. In particular, a model is only applicable in its applicability domain in terms of chemical diversity (within identified families and within a defined chemical space based on the values of descriptors in the training set) and in terms of property domain (by the range of property values in the training set). To ensure it, the following questions may be answered:

1. Is the model dedicated to the family of molecules of the target compound?
2. Is the endpoint of the model relevant regarding the application targeted for the predicted value?
3. Is the compound in the applicability domain of the model based on the molecular descriptors?
4. Is the final calculated property is the applicability domain of the model?

Moreover, the model must be applied following the exact procedure defined by the developers of the model. In particular, in the case of quantum chemical descriptors, computational descriptors, like the basis set and the method (e.g. functional), must be followed.

It is also necessary to check that the endpoint is relevant for the regulatory purpose. It is obvious when a model predicts directly a regulatory endpoint (e.g. flash point), it needs sometimes extrapolation to correlate the modeled endpoint (e.g. thermal stability) to the endpoint of regulatory interest (e.g. explosive properties). Again, no general criteria are provided to assess adequacy.

However, to prove the correct use of the QSPR model for a relevant prediction in regulatory purpose, a QSAR Prediction Reporting Format (QPRF) file has been developed as a complement to the QMRF file. This file summarizes all information needed to demonstrate that the model has been well chosen and used.

Finally, it's important to note that "Under REACH, there is no formal adoption process for (Q)SARs (or other non-testing methods) and there is no official list of accepted, legally binding models" even if harmonized documents ensuring the reproducibility of predictions and transparency in their interpretation were proposed (QMRF and QPRF files). In addition, some freely available software tools (e.g. the QSAR Toolbox) have been developed to facilitate the use of QSPR by the industry registrant for example and by regulatory bodies for regulatory purposes.

2.2 Case study 1: QSPR model dedicated to impact sensitivity with quantum descriptors

An example of validated model is the one dedicated to the impact sensitivity of nitroaliphatic compounds which is required to classify substances among explosive compounds (principle 1) (Prana et al., 2012). 34 data were used in the training set to develop the model and 16 data were kept in a validation set to test its predictivity (all data were extracted from a single reference). The final model was developed with 4 descriptors:

$$\log h_{50\%} = -0.018\ OB + 4.07\ P_{Qmax\text{-}Qmin} + 28.5\ Q^2_{NO2,max} + 4.80\ N_{O,max} - 0.438 \quad (2)$$

where $h_{50\%}$ is the impact sensitivity (in cm), OB is the oxygen balance as defined in the TDG regulation, $P_{Qmax-Qmin}$ is the polarity parameter defined by the difference between the maximum and minimum charges in the molecule, $Q^2_{NO2,max}$ is the squared of the maximum NO_2 charges obtained by natural population analysis and $N_{O,max\ is}$ the maximum nucleophilic reactivity index for a O atom (principle 2).

It presented good correlation (R^2=0.93, RMSE=0.12), robustness (Q^2LOO=0.90) and predictivity (R^2in=0.88 and RMSEin=0.19) (Principle 4). Concerning chemical interpretation (Principle 5), 3 out of 4 descriptors are quantum chemical descriptors and $N_{O,max}$ and $Q^2_{NO2,max}$ are related to the electronic and reactivity properties of nitro groups, which are critical in the decomposition process of nitro compounds. So, this model meets all five OECD principles of validation.

2.3 Case study 2: QSPR model dedicated to impact sensitivity with constitutional descriptors

If the model presented in Eq. 2 requires quantum chemical calculations, another model was developed from simpler 66 constitutional descriptors (Prana et al., 2012). It consists in a three-parameter model based on the number of nitro groups (n_{NO2}), relative number of nitrogen atoms ($n_{N,rel}$) and number of single bonds (n_{single}) (principle 2):

$$\log h_{50\%} = -2.53\ n_{N,rel} + 0.07\ n_{single} - 0.25\ n_{NO2} + 1.94 \quad (3)$$

The predicted impact sensitivity using the model as function of experimental values is plotted in the figure 2:

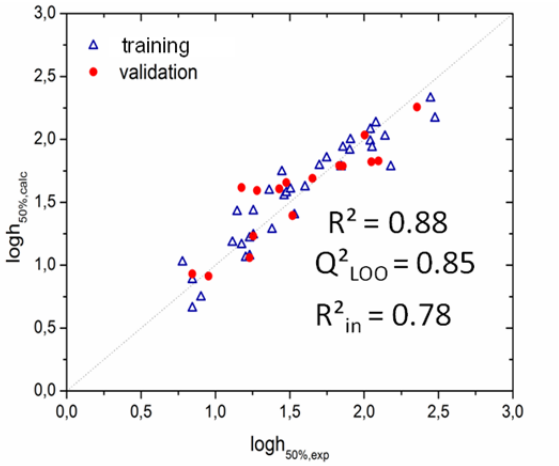

Figure 2: Experimental vs. predicted impact sensitivities of nitroaliphatics from Eq.3

This model is applicable to nitroaliphatic compounds presenting impact sensitivities between 6 and 300 cm, a relative number of nitrogen between 0.118 and 0.250, between 11 and 44 single bonds and between 2 and 12 nitro groups (Principle 3). It has been validated using a series of internal and external validation methods as summarized in Table 2 (Principle 4).

Table 2: Performances of the QSPR model of Eq.3

Training set		**Cross validation**			**Y-randomization**	
R^2	RMSE	Q^2_{LOO}	Q^2_{10CV}	Q^2_{5CV}	R^2_{YS}	SD_{YS}
0.88	0.17	0.85	0.84	0.85	0.09	0.07
Validation set						
$R^2_{EXT\ (IN)}$	$RMSE_{EXT\ (IN)}$	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	CCC	
0.81 (0.78)	0.22 (0.23)	0.81	0.81	0.83	0.93	

Concerning chemical interpretation (Principle 5), the n_{NO2} descriptor relates to the primary cleavage of the C-NO_2 bond which is known to be the main decomposition mechanism of nitroaliphatic compounds. So, this model meets all five OECD principles of validation. If this model is slightly less accurate and meaningful than the quantum chemical one, it is faster and easier to use for non-experts users as it do not need any quantum chemical calculation.

3. Conclusion

During the PREDIMOL project, INERIS identified the QSAR Toolbox as a possible way to enhance the use and (pre)-acceptability of QSPR models. As a proof of concept, the simple model developed for the impact sensitivity of nitroaliphatics presented in Eq. 3 was proposed to OECD/ECHA. After publication (Prana et al., 2012), a QMRF document was filled and sent to the specific expert comity of the QSAR Toolbox in April 2012 that stated on the scientific validity and the technical feasibility of its implementation in the platform. Finally, this model was implemented into the QSAR Toolbox in its version 3.3 in December 2014 and communication appeared in the ECHA Newsletter of February 2015. If it does not represent a legal acceptance of the model by regulatory bodies, it demonstrates the reliability of the models developed by INERIS.

Reference

Codessa software, 2002, University of Florida.

Chirico N., Gramatica, P., 2012, Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, *J. Chem. Inf. Model.*, 52 (8), 2044-2058.

Dearden J.C., Rotureau P., Fayet G., 2013, QSPR Prediction of Physico-Chemical Properties for REACH, SAR QSAR Environ. Res. 24, 279-318.

EC, Regulation (EC) N° 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH).

ECHA, European Chemicals Agency, 2008, Guidance Document on information requirements and chemical safety assessment, Chapter R.6: QSARs and grouping of chemicals, ECHA, Helsinki.

Fayet G., Rotureau P., Joubert L., Adamo C., 2011, Development of a QSPR model for predicting thermal stabilities of nitroaromatic compounds taking into account their decomposition mechanisms, J. Mol. Model., 17, 2443-2453.

Fayet G., Rotureau P., Prana V., Adamo C., 2013, Prediction of Physico-Chemical Properties for REACH Based on QSPR Models, Chemical engineering transactions, 31, 925-930, DOI: 10.3303/CET1331155.

Fayet G., Rotureau P., 2014, Development of simple QSPR models for the impact sensitivity of nitramines J. Loss Prev. Process Ind., 30, 1-8.

Gaudin T., Rotureau P., Fayet G., 2014, Combining mixing rules with QSPR models for pure chemicals to predict the flash points of binary organic liquid mixtures, Fire Safety J., 74, 61-70.

Gaudin T., Rotureau P., Fayet G., 2015, Mixture descriptors toward the development of quantitative structure-property relationship models for the flash points of organic mixtures, Ind. Eng. Chem. Res., 54, 6596-6604.

Gaussian 09, Revision B.01, 2009, Frisch, M. J. et al., Gaussian, Inc., Wallingford CT.

Gramatica P., 2007, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26, 694-701.

Jeliazkova N., Jaworska J., 2007, Ambit Discovery, version 1.20.

JRC QSAR Model Database, <http://qsardb.jrc.it> accessed 09.09.2015.

Karelson M., 2000, Molecular Descriptors in QSAR/QSPR, Wiley, New York.

Katritzky, A. R., Kuanar, M., Slavov, S., Hall, C. D., Karelson, M., Kahn, I., Dobchev, D. A., 2010, Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. Chem. Rev., 110, 5714-5789.

OECD, Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models, OECD, Paris, 2007.

Prana V., Fayet G., Rotureau P., Adamo C., Development of validated QSPR models for impact sensitivity of nitroaliphatic compounds, J. Hazard. Mater., 2012, 235-236, 169-177.

Prana V., Rotureau P., Fayet G., André D., Hub S., Vicot P., Rao L., Adamo C., Prediction of the thermal decomposition of organic peroxides by validated QSPR models, J. Hazard. Mater., 2014, 276, 216-224.

PREDIMOL project, <www.ineris.fr/predimol/> accessed 09.09.2015.

QSAR Toolbox, <http://www.qsartoolbox.org> accessed 09.09.2015.

UN, Recommendations on the transport of dangerous goods: Manual of tests and criteria, ST/SG/AC.10/Rev.5 fifth revised edition, United Nations, Geneva/New-York, 2011.

Worth A.P., in Recent Advances in QSAR Studies, 2010, eds. T. Puzyn, J. Leszczynski and M. T. Cronin, Springer Netherlands, 8, 367-382.